NATURALISTA CAMPANO
ISSN: 1827-7160
Volume 28 Issue 1, 2024

A Study on Heart Disease Prediction Using Machine Learning Algorithms

Dr. Shambhu Shankar Rai[1], Dr. Uma Durgude[2], Dr. Avinash Dhavan[3], Payal Singh[4], Dr. Sadhana Ojha[5]

[1] Assistant Professor, Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai.
[2,3,4]Assistant Professor, Bharati Vidyapeeths Institute of Management Studies and Research, Navi Mumbai.
[5] Coordinator at St. Wilfred's College of Computer Sciences.

Email: shambhumca1@gmail.com[1], dr.uma0909@gmail.com[2], avi.dhavan@gmail.com[3], payal.singh@bharatividyapeeth.edu[4], sadhanapandey459@gmail.com[5]

Abstract: The incidence of heart disease is rapidly increasing, underscoring the critical need to proactively identify potential illnesses. This diagnostic task is intricate, demanding precision and efficiency. The central focus of the research paper lies in discerning, based on diverse medical characteristics, which patients are more predisposed to heart disease. By leveraging the patient's medical history, we devised a method to assess the likelihood of a heart disease diagnosis. Employing a range of machine learning algorithms such as KNN and logistic regression, we aimed to predict and classify patients at risk of heart disease. The study also explores how the model's application can enhance the accuracy of predicting heart attacks.

Keywords: Machine Learning, Heart Disease, Prediction, Detection, Naïve Bayes.

1. Introduction

Machine learning serves as a powerful technique for extracting implicit, unknown, or known information from data, offering broad and diverse applications with expanding scope. The application of machine learning spans various classifiers, including those from supervised, unsupervised, and ensemble learning, contributing to the forecasting and accuracy determination of given datasets. The knowledge derived from machine learning can be effectively applied to projects like the Heart Disease Prediction System (HDPS), benefiting a wide range of individuals.

Cardiovascular diseases, encompassing a variety of disorders that pose potential threats to the heart, are a prevalent concern. According to the World Health Organization, there are 17.9 million deaths globally attributed to cardiovascular diseases, making it a primary contributor to adult mortality. Leveraging a person's medical history, the HDPS initiative aims to identify individuals most likely to be diagnosed with a cardiac condition. This proactive approach allows for the identification of patients exhibiting symptoms like chest pain or high blood pressure, facilitating timely and efficient diagnosis with fewer medical procedures. The ultimate goal is to streamline therapies, ensuring appropriate treatment for the patient. The study focuses on three data mining approaches, particularly emphasizing Logistic Regression and KNN, to enhance the understanding and predictive capabilities of the HDPS initiative.

The prognostication of an individual's likelihood of having heart disease has been significantly enhanced through a notably effective approach. In comparison to previously employed classifiers like naive Bayes, the suggested model demonstrates a commendable accuracy in predicting indications of heart disease in specific individuals. This achievement is attributed to the utilization of KNN and Logistic Regression. Consequently, relying on the proposed model to assess the probability of the classifier accurately and reliably identifying heart disease has alleviated considerable concerns. The system for predicting heart disease, as presented, not only enhances patient treatment but also proves cost-effective. This research yields valuable insights that can be harnessed for predicting individuals prone to developing heart disease, utilizing the .pynb file type.

Project's predictive accuracy to 87.5%. This surpasses the accuracy of the previous system, which solely relied on a single data mining technique. The inclusion of the Random Forest Classifier, along with other methods, signifies a notable improvement in predictive performance. This highlights the efficacy of employing a diverse set of data mining techniques, underscoring the potential for enhanced outcomes in predictive modeling.

Precision and efficacy in Heart Disease Prediction System (HDPS) are crucial considerations in supervised learning, which encompasses logistic regression, a technique specifically tailored for discrete values. The primary objective of this project is to ascertain the likelihood of a patient being diagnosed with cardiovascular heart

diseases based on various medical characteristics, including gender, age, chest discomfort, fasting blood sugar level, among others. The dataset, sourced from the UCI repository, comprises the patient's medical history and characteristics, serving as the basis for predicting the potential for heart disease.
In this predictive modeling endeavor, we leverage three algorithms—KNN, Random Forest Classifier, and Logistic Regression—to train on the 14 medical characteristics. Notably, KNN emerges as the most effective algorithm, achieving an accuracy rate of 88.52%. The final step involves categorizing individuals based on their risk of developing a cardiac condition, demonstrating a cost-effective and efficient procedure. This comprehensive approach aims to enhance the precision and cost-effectiveness of anticipating heart conditions in individuals through the utilization of diverse algorithms and medical characteristics.

Literature Review

Using the UCI Machine Learning dataset, extensive research has been done to predict cardiac disease. Varied data mining approaches have been used to achieve varied accuracy levels, which are detailed below.
Avinash Golande and colleagues investigate various ML algorithms that can be used to categorise cardiac disease. An investigation was conducted to examine the accuracy of the classification algorithms Decision Tree, KNN, and K-Means. The study found that Decision Trees had the highest accuracy, and it was concluded that by combining various methodologies and fine-tuning its parameters, it might be made more effective.
A system that combined the MapReduce algorithm with data mining techniques has been suggested by T. Nagamani et al. For the 45 instances in the testing set, the accuracy obtained according to this article was higher than the accuracy obtained using a traditional fuzzy artificial neural network. Here, the usage of dynamic schema and linear scaling increased the algorithm's accuracy.
AI ML model created by Fahd Saleh Alotaibi compares five alternative methods. When compared to Matlab and Weka, the Rapid Miner tool performed more accurately. This study compared the classification accuracy of Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM algorithms. The most accurate algorithm was the decision tree algorithm.
A system that employs NB (Naive Bayesian) approaches for dataset categorization and the AES (Advanced Encryption Standard) algorithm for safe data transport was proposed by Anjan Nikhil Repaka, et al.
A survey was conducted by Theresa Princy, R., et al., using various classification algorithms for heart disease prediction. The classifiers' accuracy was examined for a variety of variables using Naive Bayes, KNN (K-Nearest Neighbor), Decision Trees, and Neural Networks as the classification methodologies.
Heart disease was predicted by Nagaraj M. Lutimath et al. using Naive Bayes classification and SVM. (Support Vector Machine). Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error are the performance measurements utilized in analysis. It has been determined that SVM outperformed Naive Bayes in terms of accuracy.
After reading the aforementioned publications, the fundamental idea behind the suggested system was to build a heart disease prediction system based on the inputs presented in Table 1. By comparing the accuracy, precision, recall, and f-measure scores of the classification algorithms Decision Tree, Random Forest, Logistic Regression, and Naive Bayes, we were able to determine which classification algorithm would be most effective at predicting heart disease.
Shah et al.'s study from 2020 [18] sought to create a model for predicting cardiovascular illness using machine learning methods. The 303 cases and 17 attributes of the Cleveland heart disease dataset, which was sourced from the UCI machine learning repository, were utilised to generate the data for this project. The authors used a range of supervised classification techniques, including k-nearest neighbor, naive Bayes, decision trees, and random forests. (KKN). The study's findings showed that, at 90.8% accuracy, the KKN model had the best level of precision.
The study emphasizes the potential value of machine learning methods in anticipating cardiovascular illness and the the significance of choosing the right models and methods to get the best results.
In a study by Drod et al. (2022), the goal was to identify the most important risk factors for cardiovascular disease (CVD) in patients with metabolic-associated fatty liver disease using machine learning (ML) approaches. (MAFLD). 191 MAFLD patients had their blood biochemically analyzed, and subclinical atherosclerosis was evaluated. Using ML techniques, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis, a model to identify those with the highest risk of CVD was created. (PCA). The most important clinical traits, according to the study, were hypercholesterolemia, plaque scores, and length of diabetes. With an AUC of 0.87, the ML method worked well, correctly classifying 114/144 (79.17%) low-risk patients and 40/47 (85.11%) high-risk patients. The results of the study show that using straightforward patient criteria, an ML technique is beneficial for identifying MAFLD patients with extensive CVD.
The author of a study by Alotalibi (2019) [19] set out to look into the effectiveness of machine learning (ML) approaches for diagnosing heart failure condition. The study made use of using a dataset from the Cleveland Clinic

Foundation, we developed prediction models using a variety of ML algorithms, including decision trees, logistic regression, random forests, naive bayes, and support vector machines (SVM). During the model development process, a 10-fold cross-validation strategy was used. The findings showed that the decision tree algorithm, which had a rate of 93.19%, and the SVM method, which had a rate of 92.30%, had the highest accuracy in predicting heart disease. This work highlights the decision tree algorithm as a potential useful tool for forecasting heart failure disease and sheds light on the possibilities of ML approaches as such.

2. Methodology

This study seeks to estimate the likelihood of developing heart disease using computerised heart disease prediction, which may be useful for patients and medical professionals. We used a dataset and many machine learning methods to accomplish this goal, and the findings are presented in this study report. We intend to sanitise the data, get rid of extraneous details, and add new characteristics like MAP and BMI to improve the technique. The dataset will then be divided depending on gender, and k-modes clustering will be used. Finally, we will use the cleaned data to train the model. As shown in Figure, the revised process will result in more accurate results and greater model performance.

2.1 Data Collection and Pre-Processing

This study seeks to estimate the likelihood of developing heart disease using computerized heart disease prediction, which may be useful for patients and medical professionals.
We used a dataset and many machine learning methods to accomplish this goal, and the findings are presented in this study report. We intend to sanitize the data, get rid of extraneous details, and add new characteristics like MAP and BMI to improve the technique. The dataset will then be divided depending on gender, and k-modes clustering will be used.
Finally, we will use the cleaned data to train the model. As shown in Figure, the revised process will result in more accurate results and greater model performance.

2.2 Data Collection

The process of gathering, measuring, and analysing precise insights for study is known as data collection. A researcher can assess their hypothesis using the data that they have gathered. Regardless of the field of study, data gathering is typically the first and most crucial phase in the research process. A structured data set of Algerians who have completed analyses at the Mohand Amokrane EHS Hospital ex CNMS in Algiers, Algeria, is used in this study. It has 1200 rows and 20 columns, and the variables age, sex, cp, trestbps, chol, Ex-Ang, Col-Ves, fbs, restecg, thalach, exang, oldpeak, slope, RBP, ca, thal, smoking, alcohol use.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	29	1	1	130	204	0	0	202	0	0.0	2	0	2
2	29	1	1	130	204	0	0	202	0	0.0	2	0	2
3	29	1	1	130	204	0	0	202	0	0.0	2	0	2
4	29	1	1	130	204	0	0	202	0	0.0	2	0	2
5	34	1	3	118	182	0	0	174	0	0.0	2	0	2
6	34	0	1	118	210	0	1	192	0	0.7	2	0	2
7	34	1	3	118	182	0	0	174	0	0.0	2	0	2
8	34	0	1	118	210	0	1	192	0	0.7	2	0	2
9	34	1	3	118	182	0	0	174	0	0.0	2	0	2
10	34	0	1	118	210	0	1	192	0	0.7	2	0	2
11	35	0	0	138	183	0	1	182	0	1.4	2	0	2
12	35	1	1	122	192	0	1	174	0	0.0	2	0	2
13	35	1	0	120	198	0	1	130	1	1.6	1	0	3

Fig: Datasets for the study

2.3 Manual Investigation

Data exploration, also known as manual exploration, is the first stage of data analysis, during which users examine a sizable data collection informally to find the first patterns, traits, and areas of interest. This approach is intended

to help establish a broad picture of significant trends and key areas to investigate in more detail rather than to show every piece of information that a data set contains. To begin the pre- processing for our study, we add a column to our data collection called Target that has the values 0 or 1 (0 = not sick, 1 = sick). Figure explains this

fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	1	192	0	0.7	2	0	2	1
0	0	174	0	0.0	2	0	2	1
0	1	192	0	0.7	2	0	2	1
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0

Fig: Manual Investigation

2.4 Data Preprocessing

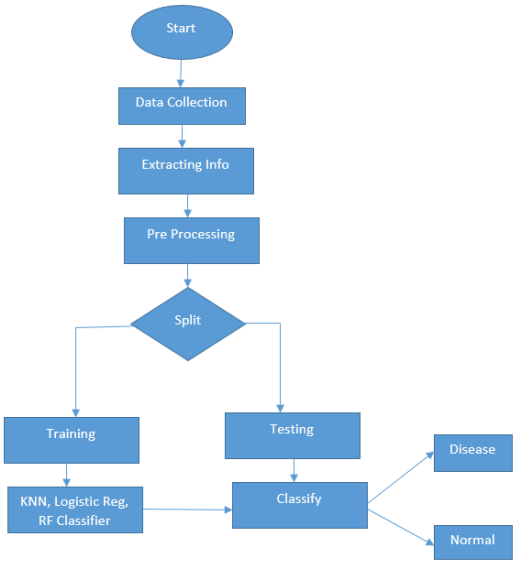
We prepare the data to be implemented before beginning the application of machine learning algorithms. This phase is accomplished in two steps:

Features choice the correlation matrix is the basis for this stage. Initially, we possessed 20 of the previously mentioned qualities. We identified 13 attributes (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) that are connected to and reliant on one another after using the Pearson correlation matrix. Table explains the specifics of the chosen characteristics.

Table: The details of selected Features

Attribute	Description	Values
Age	Age	29 to 62 years
Sex	Sex	1 - male 2 - female
CP	Chest pain type	1- typical angina pectoris 2- atypical angina 3- non-anginal pain 4- asymptomatic
trestbps	Resting blood pressure in mm/Hg	Numeric value : example: 140mm/Hg
Chol	Serum cholesterol in mg/dl	Numeric value : example: 289mg/hg
Fbs	Fasting blood pressure>120mg/dl	Numeric value : example: 129mm/Hg
Restecg	Resting electrocardiographic results	0- normal, 1- have the ST-T 2- hypertrophy
thalach	Maximum heart rate achieved	Numeric value : Example: 140,173
Exang	Exercise induced angina	1 - Yes 2 - No
Oldpeak	ST depression induced by exercise relative to rest	Numeric Value
Slope	The slope of the peak exercise ST segment	1 - on the rise 2 - flat 3- the downhill slope
Ca	Number of major vessels colored by flourosopy	0 to 3 vessels
Thal	Thalassemia	3- normal, 6- defect repaired, 7- reversible defect

3. System Implementation



3.1 Split Data Set

Dividing a data set our data set is divided into two sections: The test portion is 20% larger than the training data set in the first half. Fig. shows the division of our data set.

3.2 Testing Algorithm

We verify the accuracy of each algorithm on the various data sets after running the three algorithms on the four data sets (600, 800, 1000, and 1200 lines). Table 3 shows how we determined the accuracy for the three methods based on the confusion matrix (we used the latest data set, which had 1200 lines).
Now that we have the 4 data sets, we can show the accuracy of each algorithm, In terms of accuracy and stability against changes in the data sets, Fig. compares the three algorithms

	Neural Network Dataset (1200 lines)		SVM Dataset (1200 lines)		KNN Dataset (1200 lines)	
	Sick	Not Sick	Sick	Not Sick	Sick	Not Sick
Sick	94	8	90	11	84	18
Not Sick	6	92	9	90	11	87
Accuracy	93%		90%		85.50%	

4. Results

4.1 Shows the Risk of Heart Attack on the basis of their Age

From these findings, it is clear that even if the majority of studies use other algorithms, such as SVC and Decision trees, to identify patients with heart disease, KNN, Random Forest Classifier, and Logistic Regression produce a superior outcome to them. Our algorithms are faster and more precise than those employed by earlier studies. They also save a significant amount of money, making them very cost- effective. Furthermore, the combined maximum accuracy of KNN and Logistic Regression is 88.5%, which is higher than or nearly equal to the accuracy of earlier studies. So, to sum up that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following ‘figures’ shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

5. CONCLUSION

Heart disease has increased in prevalence throughout the world, including in our country. Consequently, diagnosing the illness before contracting it reduces the danger of dying. There has been extensive research done in this prediction field.
Our study is a component of the investigation into the identification and prognosis of cardiac disease.
It is based on the use of machine learning algorithms, of which we have selected the three most popular ones (Neural Network, SVM, and KNN), on a real data set of Algerian people, with excellent results; we reached 93% accuracy with Neural Network. The key finding of our study was that, after testing the algorithm's stability on a variety of data sets of varying sizes, it was clear that neural networks produced the greatest results. Additionally, we conducted research on feature selection and employed a correlation matrix to identify attribute dependencies. This strategy can be improved in a number of ways, including by using deep learning algorithms, other attribute selection techniques, and even bigger data sets.

6. REFERENCES

1. Golande, A., Kumar, P., "Prediction of Heart Disease Utilizing Effective Machine Learning Techniques," International Journal of Recent Technology and Engineering, Vol. 8, pp. 944-950, 2019.
2. Nagamani, T., Logeswari, S., Gomathy, B., "Heart Disease Prediction through Data Mining with Mapreduce Algorithm," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.
3. Alotaibi, F. S., "Implementing a Machine Learning Model for Heart Failure Disease Prediction," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 6, 2019.
4. Repaka, A. N., Ravikanti, S. D., Franklin, R. G., "Design and Implementation of Heart Disease Prediction Using Naive Bayesian," International Conference on Trends in Electronics and Information (ICOEI 2019).
5. Princy, T., Thomas, J., "Human Heart Disease Prediction System using Data Mining Techniques," International Conference on Circuit Power and Computing Technologies, Bangalore, 2016.
6. Lutimath, N. M., Chethan, C., Pol, B. S., "Prediction of Heart Disease using Machine Learning," International Journal of Recent Technology and Engineering, 8(2S10), pp. 474-477, 2019.
7. Ambekar, S., Phalnikar, R., "Disease Risk Prediction using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation.
8. Rjeily, C. B., Badr, G., Hassani, E., A. H., Andres, E., "Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field," in Machine Learning Paradigms, 2019, pp. 71–99.
9. Alzubi, J., Nayyar, A., Kumar, A., "Machine Learning from Theory to Algorithms: An Overview," Journal of Physics: Conference Series, 2018.
10. Alarsan, F. I., Younes, M., "Analysis and Classification of Heart Diseases using Heartbeat Features and Machine Learning Algorithms," Journal of Big Data, 2019; 6:81.